Some Thoughts on Data Stewardship – Metrics, Standards and Formats for Climate Data Records

A.M. Waple, J.J. Bates,
B.R. Nelson, D. Kim

Data and Projects at NCDC

- Remote sensing resources:

GPCP, ISCCP, POES, GOES, active & passive microwave,
NEXRAD II & III, ICOADS....etc.

- In-situ resources:

  - GHCN, USHCN, CRN, COOP etc.
  - US Palmer and other drought indices (eg. SPI)
  - Snow monitoring, extremes monitoring

Example projects:

- **North American Drought Monitor (in-situ)**
- **Air-Sea Heat Fluxes (multiple satellites and in-situ sources)**
- **ISCCP B1 data rescue (higher resolution data - reprocessing)**
- **SSTs – ERSST, blended OI**
- **Blended precip. (GPCP and in-situ)**
- **Annual State of the Climate (BAMS)**

NOAA’s Scientific Data Stewardship Program

- Overview of NOAA’s new SDS Program
- Metrics for SDS Climate Data Records
- Data Interoperability

Background to CDR Program

- In order to meet new challenges of global climate monitoring, "**creating high quality, long term datasets of global atmospheric, oceanic and terrestrial satellite observations**" is a key component of NOAA's strategy

- National Academy assisted in developing recommendations to create CDRs from satellites:

  http://www.nap.edu/html/climatedata-satellites/

- CDR: "a time series of measurements of sufficient length, consistency and continuity to determine climate variability and change"

- NOAA/NRC SDS leads

  - John J. Bates (NOAA/NCDC)

  - Mitch Goldberg (NOAA/ORA)

Key Elements of a Successful CDR Program

CDR Organizational Elements

- High-level leadership council
- Advisory council to represent climate research community and other stakeholders
- Fundamental Climate Data Record (FCDR) Teams
- Thematic Climate Data Record (TCDR) Teams

CDR Generation Elements

- High accuracy and stability of FCDRs
- Pre-launch characterization of sensors and lifetime monitoring
- Thorough calibration of sensors
- Well-defined criteria for TCDR selection
- Stakeholder involvement and feedback for TCDRs
- Well-defined criteria for TCDR validation
- Use of *in-situ* data for validation

Sustaining CDR Elements

- Available resources for reprocessing CDRs as new information becomes available
- Provisions for feedback from scientific community
- Long-term commitment of resources for generation and archiving of CDRs and associated data

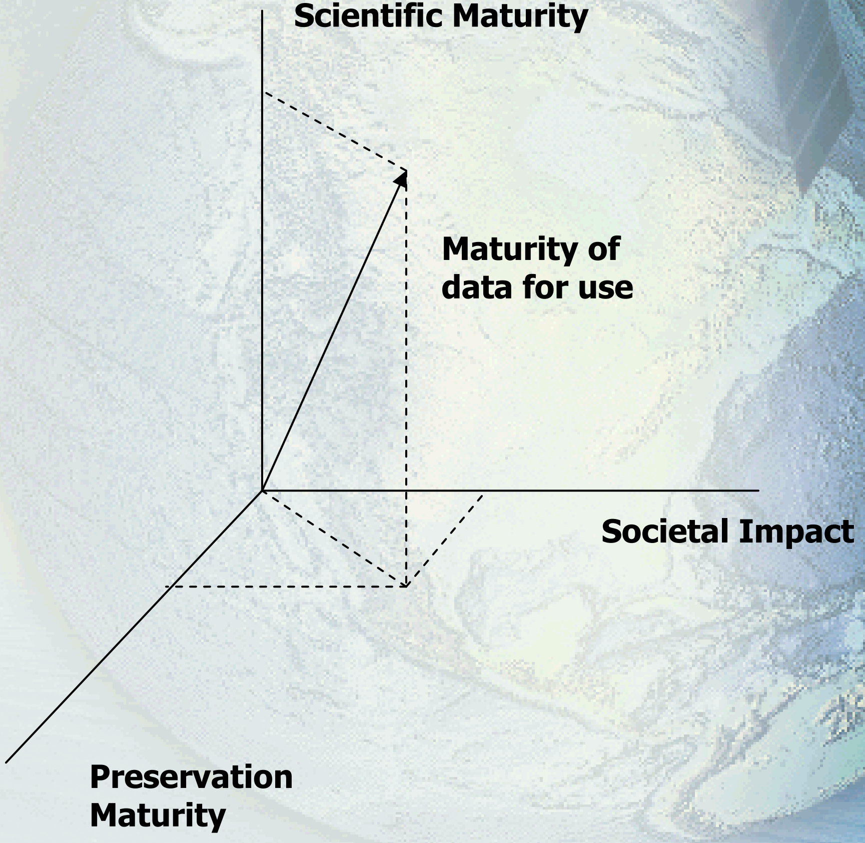
Fundamental Climate Data Record (FCDR): Time series of calibrated signals for a family of sensors together with the ancillary data used to calibrate them.

Thematic Climate Data Record (TCDR): Geophysical variables derived from FCDRs, often generated by blending satellite observations, in-situ data, and model output.

A Simple Maturity Model

Identify maturity of data products and stewardship approaches

- Represent data maturity in terms of three separate dimensions:
  - Scientific Maturity
  - Preservation Maturity
  - Societal Impact
- Total maturity is simply length of vector



Component Maturity for Climate Data Records

- Identify key attributes of maturity in each dimension
- Develop maturity ranking for each attribute on scale of 1 to 5
- Summarize component maturity by weighting each attribute
  - Simplest weight = 1/Number of attributes
  - Develop more complex weightings after experience with approach
- Advantage: can do much of work with simple spreadsheet

Key Attribute Assessment Areas –

Eg. Scientific Maturity: Public Accessibility of Data Processing

Key Assessment Area	Level 1	Level 2	Level 3	Level 4	Level 5
Number of Analysis Teams/CDR	None	Single	Two	Multiple	Consensus benchmark
Number of Independent Observing Systems/CDR	None	Single	Two	Multiple	Benchmark
Reducing Model Uncertainties: Forcings/Feedbacks/Validation	None	Single	Two	Three	Demonstrated benchmark
Availability of technique and computer code	None	Technique in one publication	Technique in multiple publications	Computer code available	Computer code available and used by other groups

Preservation Maturity Key Attributes

- Systematic Approach to Guaranteeing Preservation of Data Understanding
- Systematic Reduction of Threats to Preservation
- Assurance of Preservation Cost Effectiveness

Societal Benefit Key Attributes

- Bibliometric Metrics
  - Publications and Citations
- Scientific Community Knowledge
- Economic and Policy Utility

Some Caveats

- Using a Maturity Model will be exploratory – and iterative
  - No expectation we'll get it "right" the first time through
- Community Diversity must be incorporated
  - Different views of data processing, calibration, validation, need for knowledge preservation
  - Different vocabularies
- Deep Uncertainty needs to be incorporated
  - Diversity of opinions on areas of scientific controversy and value need common framework and disciplined discussion – openness a key
  - Including "societal benefit" is very difficult and risky

Key Benefits

- Allows us to develop an approach consistent with NRC Recommendations on Metrics
- Open Process
  - Can surface divergent needs and opinions
  - Can provide disciplined forum for discussion and resolution of differences
- Periodic Evaluation is required
  - Incorporate new information and deeper thought
  - Evaluation allows new directions

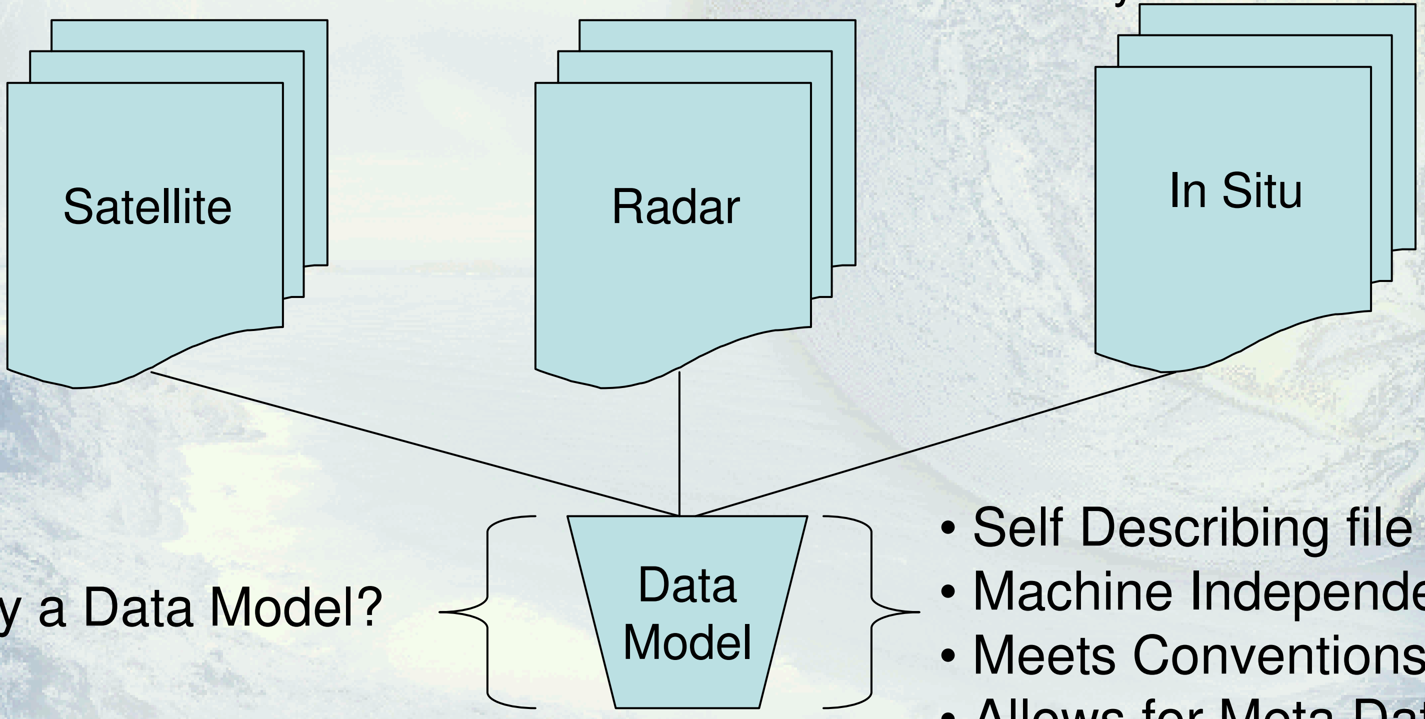
Converging on a Data Model

RSAD Data Sets

Satellite (ISCCP, AVHRR, HIRS)
Radar (NEXRAD)
In Situ (Buoy, Rain Gauge, etc.)

Data Resolution

- 2° monthly - global
- 5° monthly - global
- 20 km - global
- 2.5° monthly - 70N – 70S
- 4 km hourly - CONUS



- Why a Data Model?

- Self Describing file format
- Machine Independent
- Meets Conventions
- Allows for Meta Data

Eg. NetCDF4/HDF5

Both have wide community use and are available for a variety of operating systems

Intercompatible

•New Project underway –Beta version (November ’05), Full Release (Jan ’06)

•Users of netCDF in numerical models will benefit from support for packed data, large datasets, and parallel I/O, all of which are available with HDF5.

•Users of HDF5 will benefit from the availability of a simpler high-level interface suitable for array-oriented scientific data, wider use of the HDF5 data format, and the wealth of netCDF software for data management, analysis and visualization that has evolved among the large netCDF user community.

Conventions: COARDS/CF

- This standard is a set of conventions adopted in order to promote the interchange and sharing of files created with the netCDF Application Programmer Interface (API).

- **File Name:**

  NetCDF files should have the file name extension ".nc".
- **Coordinate Variables:**

- **Global attributes:**

:Conventions = "COARDS"; // Cooperative Ocean/Atmosphere Research Data Service

- **Data Variable attributes:**

long_name - a long descriptive name (title).
scale_factor – the data are to be multiplied by this factor
add_offset - the data are to be multiplied by this factor
missing_value - a missing value that will not be treated in any special way by the library, as the _FillValue attribute is Etc …

Conventions

Cooperative Ocean/Atmosphere Research Data Service (COARDS)
Convention for standardization of netCDF files

http://ferret.wrc.noaa.gov/noaa_coop/coop_cdf_profile.html

NetCDF Climate and Forecast (CF) Metadata Convention
Extension of COARDS

http://www.cgd.ucar.edu/cms/eaton/cf-metadata/

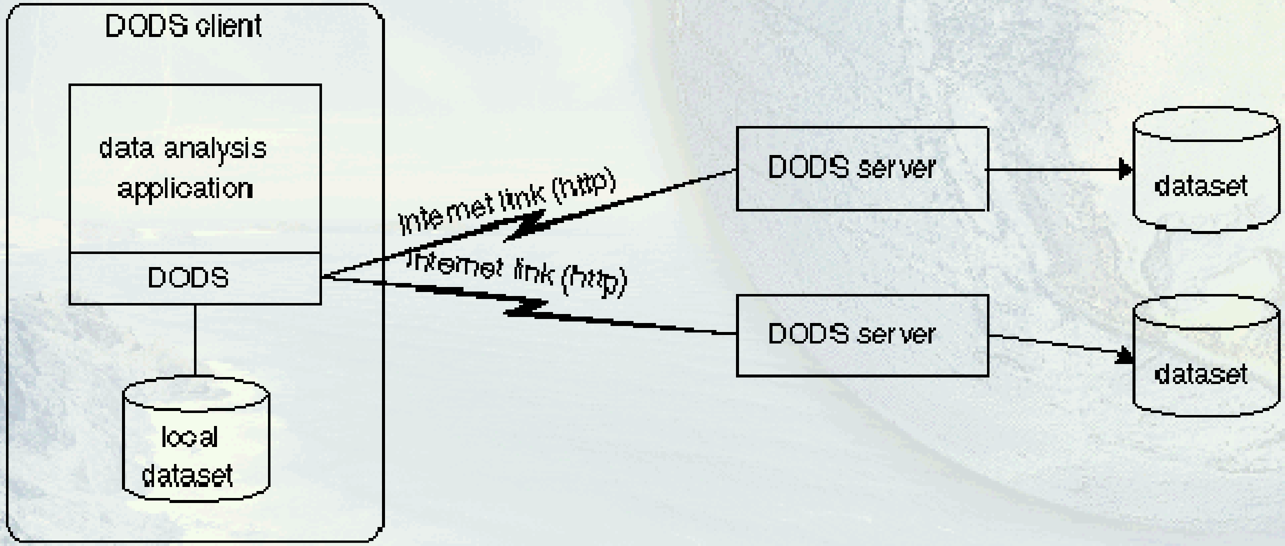
The Federal Geographic Data Committee (FGDC)

http://www.fgdc.gov/fgdc/fgdc.html

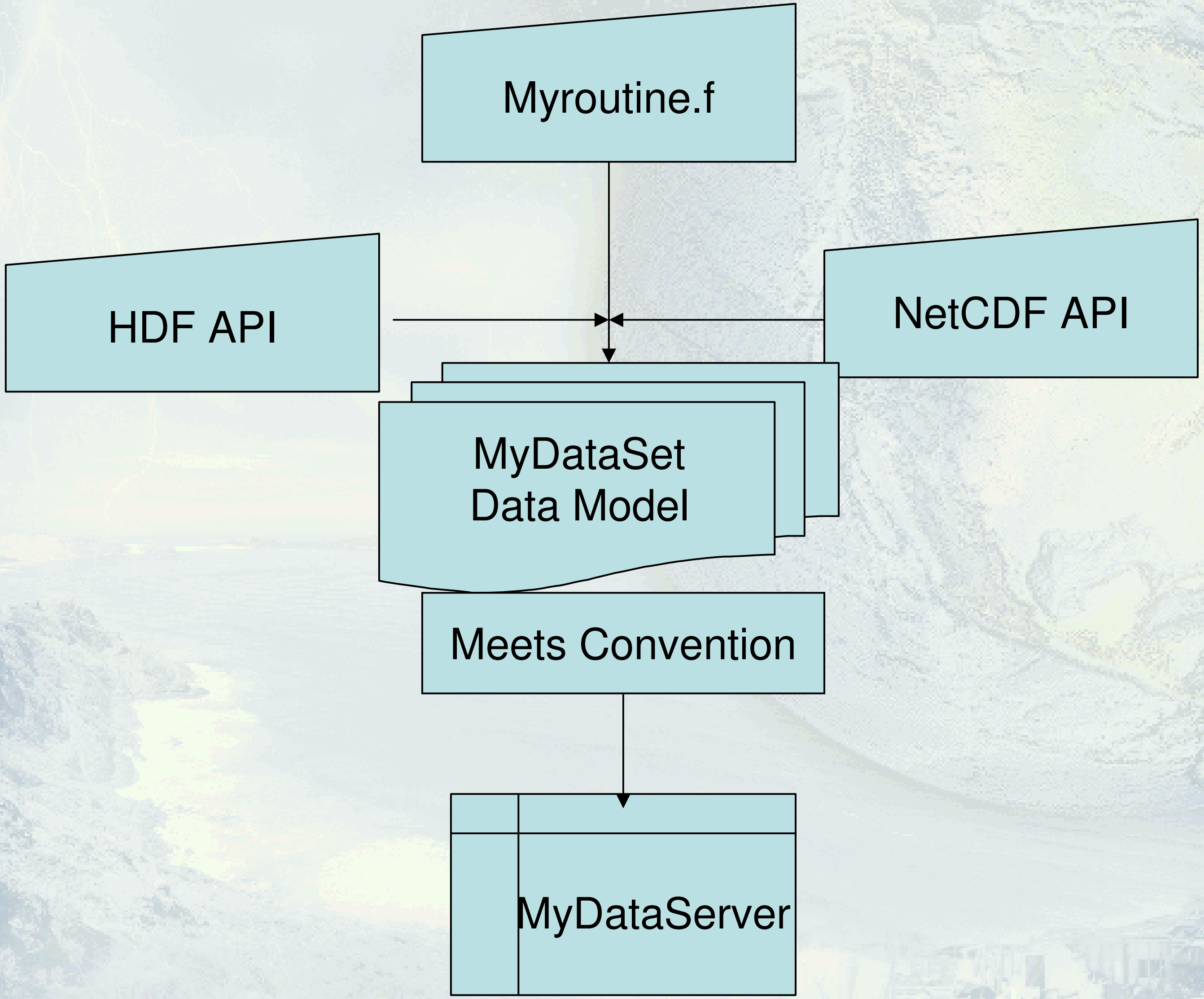
Data Access

Open source Project for a Network Data Access Protocol
(Formerly known as DODS)

- Protocol for requesting and transporting data across the Web
- Uses a Client – Server model to access data



Data Server and Data Model



Summary of Scientific Data Stewardship

- Overview of NOAA’s new SDS Program
- Metrics for SDS Climate Data Records
- Data Interoperability
- John.J.Bates@noaa.gov

SPARE SLIDES

Hierarchical Data Format

HDF4 vs HDF5

- HDF4 - Based on original 1988 version of HDF

– Backwardly compatible with all earlier versions

– 6 basic objects

- raster image, multidimensional array (SDS), palette, group (Vgroup), table (Vdata), annotation

- HDF5

– New format & library - not compatible with HDF4

– 2 basic objects

- Who uses HDF?

EOSDIS - HDF-EOS
Argonne National Laboratory
Lamont-Doherty
Los Alamos
NASA
Many others

Platforms

AIX (IBM SP)

- Cray J90, T3E
- FreeBSD
- HP-UX
- IRIX 6.5, IRIX64
- Linux
- OSF1
- Solaris
- ASCI TFLOPS
- Windows NT4.0, 98

Hierarchical Data Format

What is HDF?

- Format and software for scientific data
- Stores images, multidimensional arrays, tables, etc.
- Emphasis on storage and high performance I/O
- Free and commercial software support
- Emphasis on standards
- Users from many engineering and scientific fields

HDF5 data model

- Dataset

– multidimensional array of elements, together with supporting metadata

- Group

– directory-like structure containing datasets, groups, other objects

http://hdf.ncsa.uiuc.edu/

NetCDF Data Model

- NetCDF Data model contains dimensions, variables, and attributes

- NetCDF is a self describing file format

- NetCDF is an interface for array-oriented data access and a library that provides an implementation of the interface. The netCDF library also defines a **machine-independent** format for representing scientific data.

- http://my.unidata.ucar.edu/content/software/netcdf/index.html

- Who uses NetCDF?

  NCAR
  CDC
  PMEL
  Lamont-Doherty
  FSL
  NWS
  Many others

Platforms

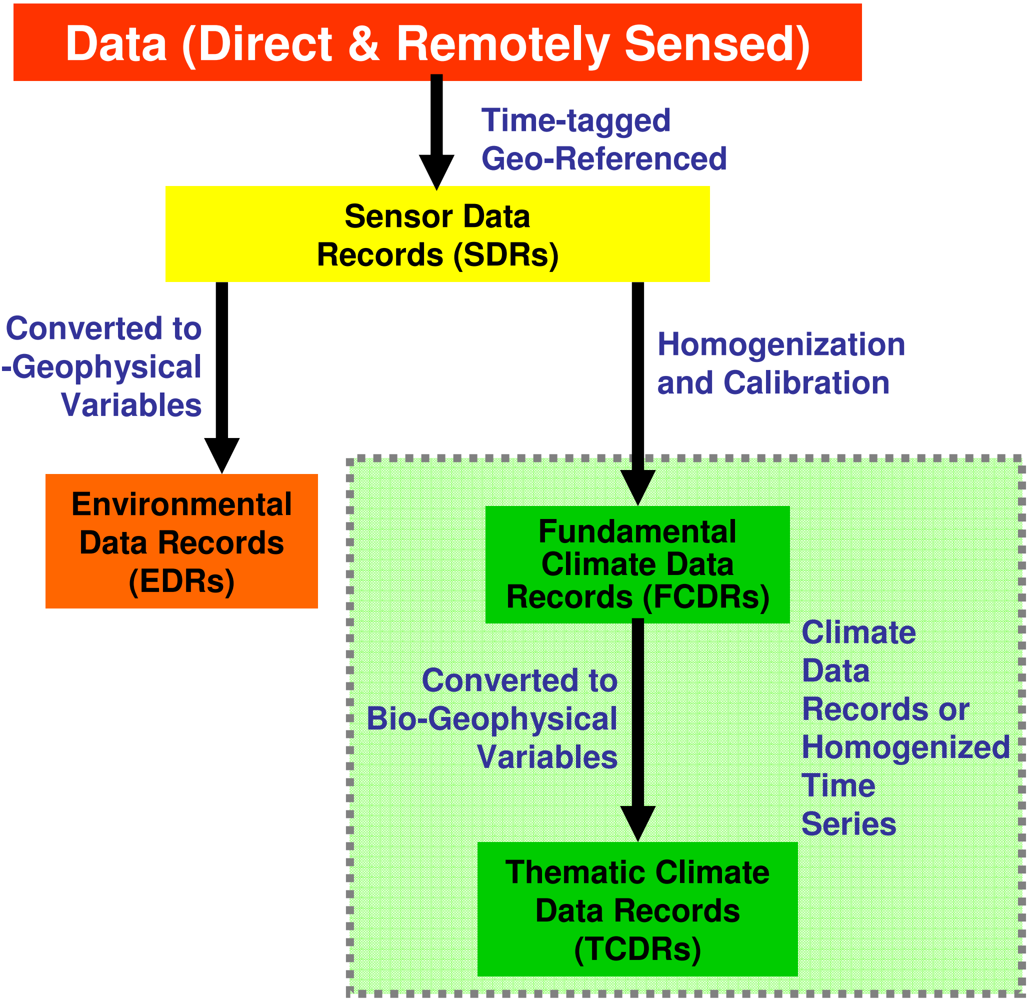
AIX-4.3
HPUX-11.00
IRIX-6.5, IRIX64-6.5
Linux 2.2
MacOS X
OSF1-4.0
SunOS-4, Solaris (Sparc and i386)
UNICOS

Software Capability-Maturity Model (CMMI) Levels

1. Initial – Unpredictable results
2. Managed – Repeatable performance
3. Defined – Cross-project interoperability
4. Quantitatively Managed – Improved performance + Compliance with Federal Enterprise Architecture
5. Optimized – Rapidly configurable performance + Continuous Process Improvement

Defining CDRs

Climate Data Records



Data (Direct & Remotely Sensed)

Time-tagged
Geo-Referenced

Sensor Data
Records (SDRs)

Converted to
Bio-Geophysical
Variables

Environmental
Data Records
(EDRs)

Homogenization
and Calibration

Fundamental
Climate Data
Records (FCDRs)

Converted to
Bio-Geophysical
Variables

Thematic Climate
Data Records
(TCDRs)

Climate
Data
Records or
Homogenized
Time
Series

Towards An Operational High Resolution Global Air-Sea Heat Fluxes

NCDC/Remote Sensing Application Div

(Huai-Min Zhang, Richard W. Reynolds, Lei Shi, John Bates)

- Overall Goal: Operational high resolution (6-hrly and 0.25º grid) heat fluxes; turbulent fluxes from multiple satellites + in-situ (NCDC); Radiation fluxes from GEWEX GRP; Combined net fluxes to be consistent with oceanic energetics.

- Present Status at NCDC: 1) Twice daily blended winds available July 1987 – present; Optimum Interpolation (OI) winds is in implementation. 2) Daily OI SST is running. 3) Ta & Qa retrieved from multiple satellites using neural network.

National Climatic Data Center

CONUS Multisensor Precipitation Estimate Reanalysis

NCDC/Remote Sensing Application Div

(Brian Nelson, Dongsoo Kim, John Bates)

- Overall Goal: High resolution precipitation analysis (hourly, 5km on NDFD grid) for NEXRAD era (1996 -) by combining NEXRAD and gauge measurement;

- Present Status at NCDC: 1) Archiving HADS historic raw gauge measurements since 1996 from NWS/OHD, 2) reprocess HADS precipitation with added QC/QA layers, 3) implemented NWS/OHD's operational Multisensor Precipitation Estimate (MPE) package.

National Climatic Data Center

ISCCP B1 Data Rescue

NCDC/Remote Sensing Application Div

(Ken Knapp, John Bates)

- Overall Goal:
  - To rescue the ISCCP B1 data for use in future cloud climate reprocessing and other climate studies
  - B1 is 10 km Geostationary IR & visible imagery from GOES, Meteosat and GMS from 1983 to present at 3 hour intervals

– Present Status at NCDC:

- ISCCP B1 data and read/navigate software are available to users
- NCDC is performing studies to assess calibration and navigation quality

National Climatic Data Center

Surface Emissivity Database Derived from DMSP/SSMI and ISCCP B1 datasets

NCDC/Remote Sensing & Applications Division

(Hilawe Semunegus, Kenneth Knapp, John Bates)

- Overall Goal: Collocation of DMSP/SSMI (dual-polarized brightness temperatures at 19 Ghz, 22 Ghz (vertical only), 37 Ghz, and 85 Ghz) and cloud-cleared ISCCP B1 IR brightness temperatures for the estimation of global land-surface emissivities.

- Present Status and Near Future plan at NCDC: 1) final stages of developing a robust matching algorithm for the two datasets 2) establishing regional and temporal boundaries for a test case. 3) create a database that will be accessible to users online.

National Climatic Data Center